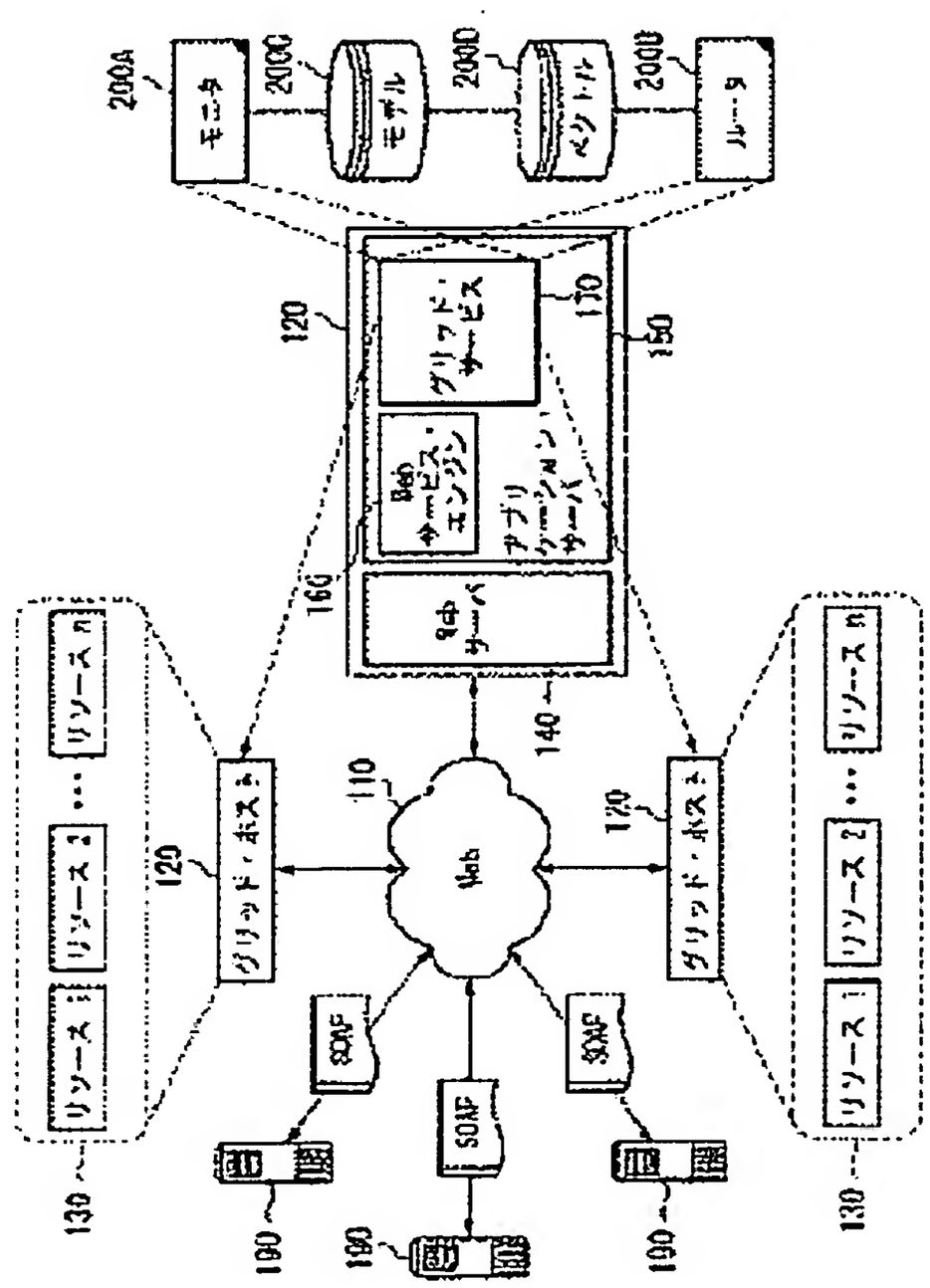
AUTONOMIC SERVICE ROUTING USING OBSERVED RESOURCE REQUIREMENT FOR SELF-OPTIMIZATION

Publication number: JP2004252975
Publication date: 2004-09-09
Inventor: DOYLE RONALD P; KAMINSKY DAVID LOUIS
Applicant: IBM
Classification:
- international: ***G06F15/177; G06F9/46; G06F9/50; G06F15/16; H04L12/56; H04L29/00;*** **G06F9/46; G06F15/16; H04L12/56; H04L29/00;** (IPC1-7): G06F15/177; G06F9/46; G06F15/16; H04L12/56
- European: G06F9/46A4
Application number: JP20040032742 20040209
Priority number(s): US20030370837 20030221

Also published as:
US2004167959 (A1)
CN1523845 (A)

Abstract of **JP2004252975**

PROBLEM TO BE SOLVED: To provide a system and a method that route a service request in the field of distributed computing including Web services and grid services.
SOLUTION: The system can include a model table configured to store resource models. A monitor can be coupled to the model table, and programmed both to model resource consumption in a service providing infrastructure and to store the modeled resource consumption in the model table. A router also can be coupled to the model table. Specifically, the router can be programmed to route each service request to a corresponding service instance disposed in an associated service host having a service providing infrastructure. In a preferred mode, the associated service host can include a grid host in a grid computing system.
COPYRIGHT: (C)2004,JPO&NCIPI



PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2004-252975
(43)Date of publication of application : 09.09.2004

(51)Int.Cl. G06F 15/177
G06F 9/46
G06F 15/16
H04L 12/56

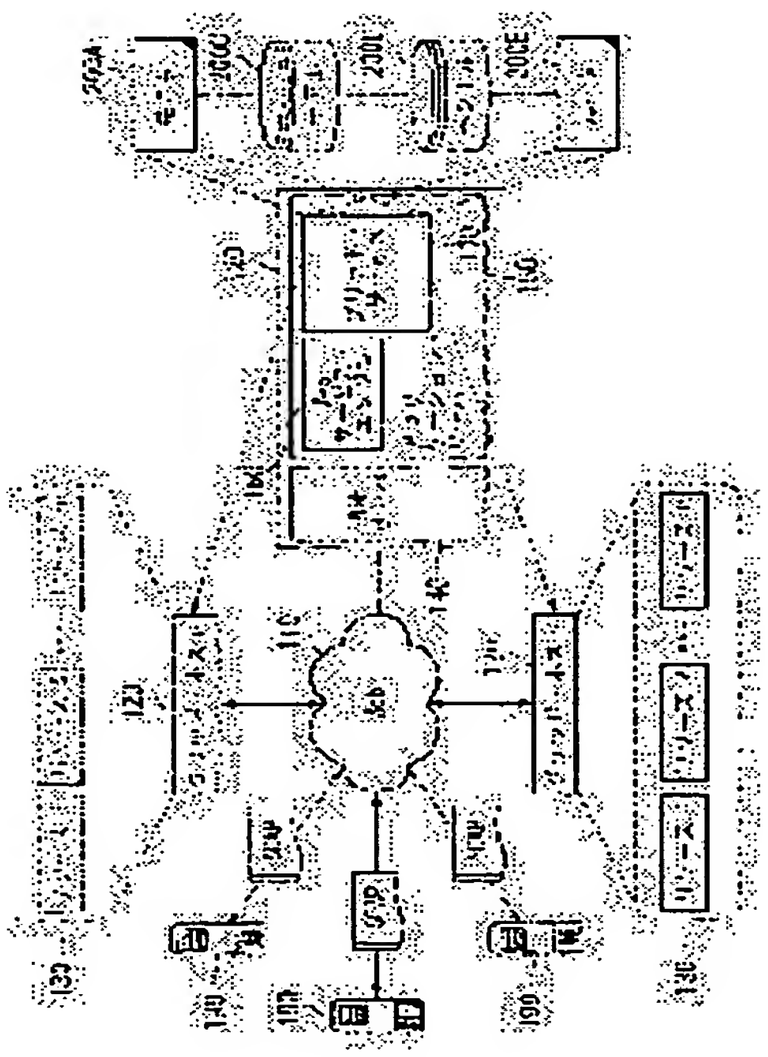
(21)Application number : 2004-032742
(22)Date of filing : 09.02.2004

(71)Applicant : INTERNATL BUSINESS MACH CORP <IBM>
(72)Inventor : DOYLE RONALD P
KAMINSKY DAVID LOUIS

(30)Priority
Priority number : 2003 370837 Priority date : 21.02.2003 Priority country : US

(54) AUTONOMIC SERVICE ROUTING USING OBSERVED RESOURCE REQUIREMENT FOR SELF-OPTIMIZATION

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a system and a method that route a service request in the field of distributed computing including Web services and grid services.
SOLUTION: The system can include a model table configured to store resource models. A monitor can be coupled to the model table, and programmed both to model resource consumption in a service providing infrastructure and to store the modeled resource consumption in the model table. A router also can be coupled to the model table. Specifically, the router can be programmed to route each service request to a corresponding service instance disposed in an associated service host having a service providing infrastructure. In a preferred mode, the associated service host can include a grid host in a grid computing system.



LEGAL STATUS

[Date of request for examination] 09.02.2004
[Date of sending the examiner's decision of rejection] 17.10.2006
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁(JP)

(12)公開特許公報(A)

(11)特許出願公開番号
特開2004-252975
(P2004-252975A)

(43)公開日 平成16年9月9日(2004.9.9)

(51)Int.Cl.[7]		F I			テーマコード（参考）
G06F	15/177	G06F	15/177	674A	5B045
G06F	9/46	G06F	9/46	360C	5B098
G06F	15/16	G06F	15/16	620B	5K030
H04L	12/56	H04L	12/56	100Z	

審査請求 有　請求項の数 16　OL　（全 13 頁）

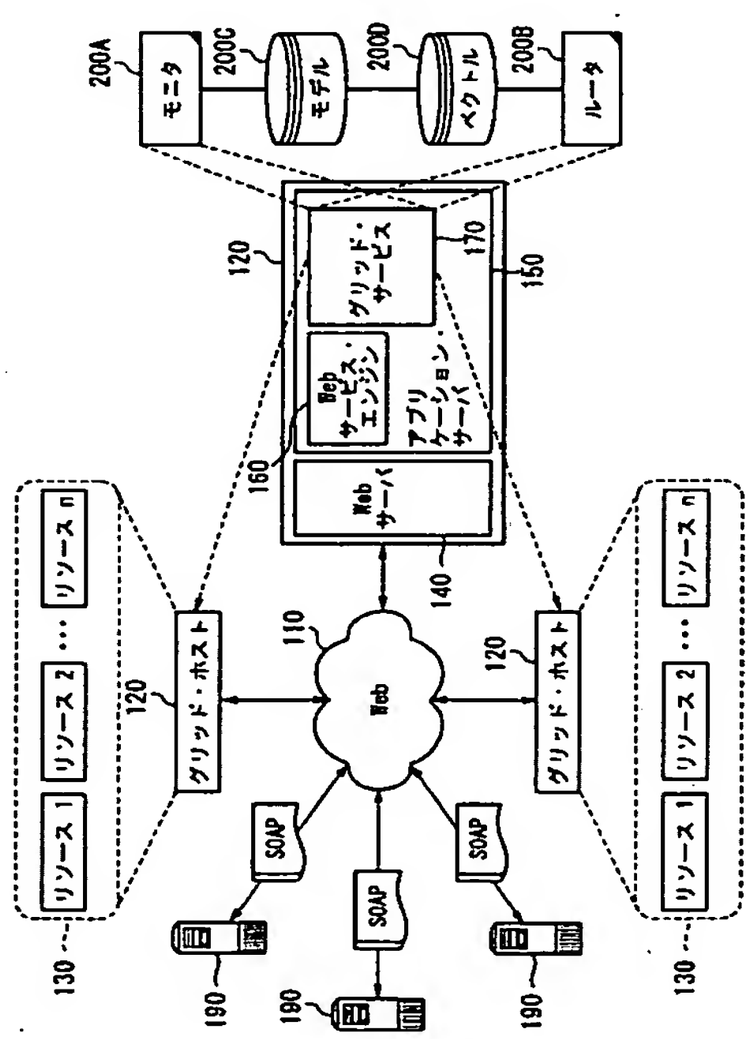
(21)出願番号　特願2004-32742 (P2004-32742)
(22)出願日　平成16年2月9日 (2004.2.9)
(31)優先権主張番号　10/370837
(32)優先日　平成15年2月21日 (2003.2.21)
(33)優先権主張国　米国 (US)

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
INTERNATIONAL BUSINESS MASCHINES CORPORATION
アメリカ合衆国10504　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
(74)代理人　100086243
弁理士　坂口　博
(74)代理人　100091568
弁理士　市位　嘉宏
(74)代理人　100108501
弁理士　上野　剛史

最終頁に続く

(54)【発明の名称】自己最適化のために観察されたリソース要件を使用するオートノミック・サービス・ルーティング

(57)【要約】
【課題】　サービス要求をルーティングするシステムおよび方法を提供すること。
【解決手段】　本システムは、リソース・モデルをストアするように構成されたモデル・テーブルを含むことができる。モニタをモデル・テーブルに結合することができ、サービス提供インフラストラクチャ内のリソース消費をモデル化するように、およびモデル化されたリソース消費をモデル・テーブル内にストアするようにもプログラミングすることができる。モデル・テーブルには、ルータも結合することができる。特に、ルータは、各サービス要求を、サービス提供インフラストラクチャを有する関連付けられたサービス・ホスト内に配置された対応するサービス・インスタンスにルーティングするようにプログラミングすることができる。本発明の好ましい態様では、関連付けられたサービス・ホストは、グリッド・コンピューティング・システムにグリッド・ホストを含めることができる。
【選択図】　図1



【特許請求の範囲】
【請求項１】
サービス要求ルーティング・システムであって、
リソース・モデルをストアするように構成されたモデル・テーブルと、
前記モデル・テーブルに結合され、さらに、サービス提供インフラストラクチャ内のリソース消費をモデル化するように、および前記モデル化されたリソース消費を前記モデル・テーブル内にストアするようにもプログラミングされたモニタと、
前記モデル・テーブルに結合され、各サービス要求を、前記サービス提供インフラストラクチャのリソース構成要素と前記サービス要求のリソース・モデルとの突合せに基づいて、サービス提供インフラストラクチャを有する関連付けられたサービス・ホスト内に配置された対応するサービス・インスタンスにルーティングするようにプログラミングされたルータと
を含む、サービス要求ルーティング・システム。
【請求項２】
前記関連付けられたサービス・ホストはグリッド・コンピューティング・システム内にグリッド・ホストを含む、請求項１に記載のサービス要求ルーティング・システム。
【請求項３】
前記モデル・テーブル内の各リソース・モデルは時系列モデルである、請求項１に記載のサービス要求ルーティング・システム。
【請求項４】
前記リソース構成要素は前記サービス提供インフラストラクチャに対応するリソース・ベクトルを形成する、請求項１に記載のサービス要求ルーティング・システム。
【請求項５】
前記リソース・ベクトル内の各リソース構成要素は、サーバ・タイプ、帯域幅、およびストレージシステム・タイプからなるグループから選択されたリソースを含む、請求項４に記載のサービス要求ルーティング・システム。
【請求項６】
個々のリソース・ベクトル間の相対的コストを決定するために、各リソース・ベクトルのスカラ・コストを比較するようにプログラミングされたコンパレータをさらに含む、請求項４に記載のサービス要求ルーティング・システム。
【請求項７】
サービス要求をサービス提供インフラストラクチャのサービス・インスタンスにルーティングするための方法であって、
サービス要求を受け取るステップと、
対応するサービス提供インフラストラクチャを有する少なくとも２つのサービス・ホストについて、リソース・ベクトルを計算するステップと、
前記サービス要求についてリソース・モデルを検索するステップと、
最適のリソース・ベクトルを識別するために、前記検索されたリソース・モデルを前記リソース・ベクトルそれぞれと突き合せるステップと、
前記サービス要求を、前記識別された最適のリソース・ベクトルに関連付けられた選択されたサービス・ホストにルーティングするステップと
を含む、方法。
【請求項８】
前記計算するステップは、前記リソース・ベクトルのそれぞれについて、サーバ・タイプ、サーバ・パフォーマンス、サーバ容量、処理帯域幅、通信帯域幅、ストレージ・タイプ、ストレージ容量、およびストレージ・パフォーマンスからなるグループから選択された、少なくとも２つのスカラ・リソース構成要素を計算するステップを含む、請求項７に記載の方法。
【請求項９】
前記計算するステップは、前記リソース・ベクトルのそれぞれについてスカラ・コストを

計算するステップをさらに含む、請求項７に記載の方法。
【請求項１０】
受け取ったサービス要求の処理をモニタするステップと、
前記処理中に消費される、サービス・ホスト内の個々のリソース構成要素を識別するステップと、
前記サービス・ホスト内の前記識別された個々のリソース構成要素に基づいて、前記サービス要求についてのリソース・モデルを生成するステップと
をさらに含む、請求項７に記載の方法。
【請求項１１】
前記生成するステップは、前記識別された個々のリソース構成要素に基づいて、前記サービス要求のそれぞれについて時系列モデルを計算するステップを含む、請求項１０に記載の方法。
【請求項１２】
サービス要求をサービス提供インフラストラクチャのサービス・インスタンスにルーティングするために、コンピュータ・プログラムをストアしたマシン読取り可能ストレージであって、前記コンピュータ・プログラムは、
サービス要求を受け取るステップと、
対応するサービス提供インフラストラクチャを有する少なくとも２つのサービス・ホストについて、リソース・ベクトルを計算するステップと、
前記サービス要求についてリソース・モデルを検索するステップと、
最適のリソース・ベクトルを識別するために、前記検索されたリソース・モデルを前記リソース・ベクトルのそれぞれと突き合せるステップと、
前記サービス要求を、前記識別された最適のリソース・ベクトルに関連付けられた選択されたサービス・ホストにルーティングするステップと
を前記マシンに実行させるための命令のルーチン・セットを含む、マシン読取り可能記憶装置。
【請求項１３】
前記計算するステップは、前記リソース・ベクトルのそれぞれについて、サーバ・タイプ、サーバ・パフォーマンス、サーバ容量、処理帯域幅、通信帯域幅、ストレージ・タイプ、ストレージ容量、およびストレージ・パフォーマンスからなるグループから選択された、少なくとも２つのスカラ・リソース構成要素を計算するステップを含む、請求項１２に記載のマシン読取り可能記憶装置。
【請求項１４】
前記計算するステップは、前記リソース・ベクトルのそれぞれについてスカラ・コストを計算するステップをさらに含む、請求項１２に記載のマシン読取り可能記憶装置。
【請求項１５】
受け取ったサービス要求の処理をモニタするステップと、
前記処理中に消費される、サービス・ホスト内の個々のリソース構成要素を識別するステップと、
前記サービス・ホスト内の前記識別された個々のリソース構成要素に基づいて、前記サービス要求についてのリソース・モデルを生成するステップと
をさらに含む、請求項１２に記載のマシン読取り可能記憶装置。
【請求項１６】
前記生成するステップは、前記識別された個々のリソース構成要素に基づいて、前記サービス要求のそれぞれについて時系列モデルを計算するステップを含む、請求項１５に記載のマシン読取り可能記憶装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、Ｗｅｂサービスおよびグリッド・サービスを含む分散コンピューティングの

分野に関し、より詳細には、サービス要求をサービス提供インフラストラクチャ内のサービス・インスタンスにルーティングすることに関する。

【背景技術】

【０００２】

Ｗｅｂサービスは分散コンピューティングの最先端であり、ワールド・ワイド・ウェブを介した構成要素ベース・アプリケーションの急速な開発をサポートするために真のユニバーサル・モデルを開発するための基礎とみなされている。Ｗｅｂサービスは、当分野では、サービス指向の構成要素ベース・アプリケーション・アーキテクチャを記した数多くの新標準を含むものとして知られている。特に、Ｗｅｂサービスは緩やかに結合された、語義的には個別の機能をカプセル化した再使用可能なソフトウェア構成要素であり、分散されて、標準的なインターネット・プロトコルを介してプログラムに従ってアクセスすることができる。

【０００３】

概念的に言えば、Ｗｅｂサービスとは、プロセス内の離散タスクが価値ネット全体にわたって広く分散されたモデルを表現する。とりわけ、多くの業界エキスパート達は、サービス指向のＷｅｂサービスがインターネットの次の発展段階においてイニシアティブを握るものと考えている。通常、Ｗｅｂサービスはウｅｂサービス定義言語（ＷＳＤＬ）などのインターフェースによって定義付けることが可能であり、インターフェースに従って実施可能であるが、実施がＷｅｂサービス・インターフェースに準拠している限り、その細部は問題ではない。Ｗｅｂサービスが対応するインターフェースに従って実施されると、その実施は、当分野でよく知られたUniversal Description, Discover and Integration（ＵＤＤＩ）などのＷｅｂサービス・レジストリに登録することができる。登録すると、サービスの要求者は、たとえば、Simple Object Access Protocol（ＳＯＡＰ）を含む、任意のサポーティング・メッセージング・プロトコルを使用することで、Ｗｅｂサービスにアクセスすることができる。

【０００４】

Ｗｅｂサービスをサポートするサービス指向アプリケーション環境では、信頼できるサービスを見つけ、それら信頼できるサービスをアプリケーションの目的に合わせてリアルタイムで動的に統合することには問題があることがわかっている。レジストリ、ディレクトリ、およびディスカバリ・プロトコルは、サービスの検出およびサービス対サービスの相互接続論理を実施するための基本構造を提供するものであり、レジストリ、ディレクトリ、およびディスカバリ・プロトコル単独では、分散相互運用性には適さない。むしろ、統一アプリケーションの形成においてＷｅｂサービスの分散を容易にするためには、より構造化され、形式化されたメカニズムが必要である。

【０００５】

とりわけ、Open Grid ServicesArchitecture（ＯＧＳＡ）を介したグリッド・メカニズムの生理機能は、さもなければレジストリ、ディレクトリおよびディスカバリ・プロトコルの排他的使用では不可能な方法で、以下では「グリッド・サービス」と呼ぶ分散されたシステムにまたがるＷｅｂサービスの発見と結合の両方において、プロトコルを提供することができる。Ian Foster, Carl Kesselman, およびSteven Tueckeによる”The Anatomy of the Grid”, Intl J. Supercomputer Applications、２００１年、およびIan Foster, Carl Kesselman, Jeffrey M. NickおよびSteven Tueckeによる”The Physiology of the Grid”, Globus.org、２００２年６月２２日の両方に記載されるように、グリッド・メカニズムは、分散コンピューティング・インフラストラクチャを提供することができ、それを介して要求側クライアントがグリッド・サービス・インスタンスを作成、命名、および発見できる。

【０００６】

グリッド・サービスは、拡張リソースの共用およびスケジューリングのサポート、高度な分散アプリケーションに共通して必要な長寿命状態のサポート、ならびに企業間コラボレーションのサポートを提供することにより、単なるＷｅｂサービスを延長するものであ

る。さらに、Ｗｅｂサービスは永続的なサービスの発見および呼出しに対処するだけのものである一方、グリッド・サービスは、動的に作成および破壊が可能な一時的サービス・インスタンスをサポートするものである。グリッド・サービスを使用する著しい利点には、コンピューティング・リソースのより効率的な使用による情報技術の所有コストの削減、および様々なコンピューティング構成要素の統合の容易性の向上が含まれる。したがって、グリッド・メカニズム、詳細には、ＯＧＳＡに準拠したグリッド・メカニズムは、サービス指向アーキテクチャを実装することが可能であり、それによって、たとえ組織の領域をまたがる場合であっても分散システム統合の基礎を提供することができる。

【０００７】

コンピューティング・グリッド内では、サービス提供インフラストラクチャは、グリッド・サービスなどの分散サービスの実行をホストするための処理リソースを提供することができる。サービス提供インフラストラクチャは、サーバ・コンピューティング・デバイスと、直接接続ストレージ、ネットワーク接続ストレージ、およびストレージ・エリア・ネットワークを含むストレージ・システム、処理と、および通信の帯域幅などを含む、リソース・セットを含むことができる。サービス提供インフラストラクチャ内で処理される個々のトランザクションは、これらリソースの様々な混合を消費することができる。

【０００８】

グリッド・サービスのコンテキストでは、指定されたサービス提供インフラストラクチャ内でホストされる特定のサービス・インスタンスに、特定サービス・インスタンスの待ち行列長さに従って、要求をルーティングすることが知られている。待ち行列長さに基づいた特定のサービス・インスタンスの論理選択は、サービス処理の要求を可能な最短の待ち行列に配置することによって、応答時間を最小にするための試みを表す。同様に、ホスティングサービス提供インフラストラクチャの処理機能を、特定のサービス・インスタンスを選択する際に考慮に入れることができる。

【０００９】

より詳細には、他のサービス・インスタンスの２倍の速さで要求を処理できる特定サービス・インスタンスは、特定サービス・インスタンスが他のサービス・インスタンスの待ち行列の２倍の待ち行列を有する場合、他のサービス・インスタンスと同じ処理スループットを有することができる。さらに、待ち行列長さ選択ストラテジは細分性が粗すぎる可能性があり、要求されたサービスのリソース要件をサービス提供インフラストラクチャの使用可能なリソースと突き合わせるものではない。特に、従来の環境では、単なるスカラ・ベンチマークだけをサービス提供インフラストラクチャ全体に関連付けることができる。したがって、サービス提供インフラストラクチャの細分性の高い構成要素を、考慮に入れることは決してない。

【非特許文献１】Ian Foster,Carl Kesselman, Steven Tuecke, ”The Anatomy of the Grid”, Intl J. SupercomputerApplications (2001)

【非特許文献２】IanFoster, Carl Kesselman, Jeffrey M. Nick, Steven Tuecke, “The Physiology of theGrid”, Globus. Org (June 22, 2002)

【非特許文献３】” An Open GridServices Architecture”, Globus Tutorial, Argonne National Labatory, (January 29,2002)

【発明の開示】

【発明が解決しようとする課題】

【００１０】

本発明の目的は、サービス要求をルーティングするシステムおよび方法を提供することである。

【課題を解決するための手段】

【００１１】

本発明によれば、個々のサービス要求を、サービス要求のリソース要件および消費パターンに最も適合したリソース構成要素を有する、選択されたサービス・ホスト内のサービス・インスタンスにルーティングすることができる。このようにして、単なるスカラ・ベ

ンチマークだけをサービス提供インフラストラクチャ全体に関連付けることのできる従来の環境とは異なり、サービス要求をサービス・インスタンスにルーティングするときに、グリッド・ホストのサービス提供インフラストラクチャの細分性の高い構成要素を考慮に入れることができる。

【００１２】

サービス要求ルーティング・システムは、リソース・モデルをストアするように構成されたモデル・テーブルを含むことができる。モニタをモデル・テーブルに結合することが可能であり、サービス提供インフラストラクチャ内のリソース消費をモデル化するように、およびモデル化されたリソース消費をモデル・テーブル内にストアするようにもプログラミングすることができる。モデル・テーブルには、ルータも結合することができる。。特に、ルータは、各サービス要求を、サービス提供インフラストラクチャを有する関連付けられたサービス・ホスト内に配置された対応するサービス・インスタンスにルーティングするようにプログラミングすることができる。本発明の好ましい態様では、関連付けられたサービス・ホストは、グリッド・コンピューティング・システムにグリッド・ホストを含めることができる。

【００１３】

重要なことに、ルーティングは、サービス提供インフラストラクチャのリソース構成要素とサービス要求のリソース・モデルとの突き合わせに基づくことができる。さらに、好ましい態様では、モデル・テーブル内の各リソース・モデルは時系列モデルであってよい。最終的に、リソース構成要素は、サービス提供インフラストラクチャに対応するリソース・ベクトルを形成することができる。この点に関して、リソース・ベクトル内の各リソース構成要素は、サーバ・タイプ、帯域幅、およびストレージ・システム・タイプからなるグループから選択されたリソースを含むことができる。他のリソースは、たとえば、キャッシュ・サイズまたはＣＰＵ速度などのより細分性の高いコンピューティング・リソースを含むことができる。さらに、個々のリソース・ベクトル間の相対的コストを決定するために、各リソース・ベクトルのスカラ・コストを比較するようにプログラム可能な、コンパレータを含むこともできる。

【００１４】

サービス要求をサービス提供インフラストラクチャのサービス・インスタンスにルーティングする方法は、サービス要求を受け取ること、および少なくとも２つのサービス・ホストについてリソース・ベクトルを計算することを含むことができる。各サービス・ホストは、対応するサービス提供インフラストラクチャを有することができる。サービス要求について、リソース・モデルを検索することができる。したがって、最適のリソース・ベクトルを識別するために、検索されたリソース・モデルをリソース・ベクトルのそれぞれと突き合せることができる。最終的に、サービス要求を、識別された最適のリソース・ベクトルに関連付けられた選択されたサービス・ホストにルーティングすることができる。

【００１５】

本発明の好ましい態様では、リソース・ベクトルのそれぞれについて、少なくとも２つのスカラ・リソース構成要素を計算することができる。この点に関して、スカラ構成要素は、サーバ・タイプ、サーバ・パフォーマンス、サーバ容量、処理帯域幅、通信帯域幅、ストレージ・タイプ、ストレージ容量、およびストレージ・パフォーマンスを含むことができる。また、リソース・ベクトルのそれぞれについて、スカラ・コストを計算することもできる。このようにして、スカラ・コストを比較し、よりコスト効果の高いリソース・ベクトルを決定することができる。

【００１６】

リソース・モデルを生成するために、受け取ったサービス要求の処理をモニタすることが可能であり、処理中に消費されるサービス・ホスト内の個々のリソース構成要素を識別することができる。その結果、サービス・ホスト内の識別された個々のリソース構成要素に基づいて、サービス要求についてのリソース・モデルを生成することができる。特に、この生成ステップは、識別された個々のリソース構成要素に基づいて、サービス要求のそ

れぞれについて時系列モデルを計算するステップを含むことができる。

【００１７】

現在の好ましい実施形態が図面に示されているが、本発明は、図示された精密な配置構成および手段に限定されるものでないことを理解されたい。

【発明を実施するための最良の形態】

【００１８】

本発明は、選択されたサービス提供インフラストラクチャ内のサービス・インスタンスにサービス要求をルーティングするための方法およびシステムである。特に、サービスの細分性の高いリソース要件を、要求されたサービスのインスタンスをホスティングするサービス提供インフラストラクチャに関連付けられたリソース・セット内にある細分性の高いリソースと突き合せることができる。要求されたサービスのリソース要件と、ホストのサービス提供インフラストラクチャのリソース可用性との最適の突合せに基づいて、サービス処理の要求を、適合したサービス提供インフラストラクチャ内でホストされるサービス・インスタンスに割り当てることができる。このようにして、サービス提供インフラストラクチャの細分性の粗いスカラ評価に基づいた、サービス要求の単なるルーティングを回避することができる。

【００１９】

図１は、本発明に従い、要求されたサービスのリソース要件に最も適したリソースを有するサービス提供インフラストラクチャ内でホストされるサービス・インスタンスに、サービス要求をルーティングするように構成されたサービス・グリッドのブロック図である。当分野の技術者であれば明らかなように、サービス・グリッドは、たとえばインターネットなどのコンピュータ通信ネットワーク１１０を横切ってグリッド様式で相互に通信可能なようにリンクされた、１つまたは複数のグリッド・ホスト１２０で構成されたＷｅｂサービス・グリッドとすることができる。個々の要求側クライアント１９０は、１つまたは複数のグリッド・ホスト１２０にＷｅｂサービスへのアクセスを要求することができる。特に、当分野で周知であるように、要求側クライアント１９０とグリッド・ホスト１２０との間で、ＳＯＡＰ符号化メッセージを交換することができる。メッセージは、特定のＷｅｂサービスの位置を発見するための要求、ならびに要求されたＷｅｂサービスのネットワーク位置が明示された要求に対する応答を含むことができる。

【００２０】

グリッド・ホスト１２０は、中央集中方式で１つのサーバ・コンピューティング・デバイス内に配置するか、または分散方式で複数のサーバ・コンピューティング・デバイスにまたがって配置することができる。どちらの場合も、マークアップ文書などのコンテンツに関するネットワーク要求に応答するように構成可能な、Ｗｅｂサーバ１４０を提供することができる。当分野の通常の技術者であれば理解されるように、Ｗｅｂサーバ１４０は、ハイパーテキスト転送プロトコル（ＨＴＴＰ）メッセージを処理するために、およびハイパーテキスト・マークアップ言語（ＨＴＭＬ）形式文書、拡張可能マークアップ言語（ＸＭＬ）形式文書などのマークアップを配布するように構成することができる。

【００２１】

Ｗｅｂサーバ１４０は、グリッド・ホスト１２０内でアプリケーション・サーバ１５０と通信可能なようにリンクすることができる。アプリケーション・サーバは当分野で周知であり、通常は、インタープリタ方式またはネイティブ形式のいずれかでマシン・コードを処理するように構成される。従来のアプリケーション・サーバは、スクリプトおよびサーブレットなどのサーバ側論理を処理するものである。いかなる場合でも、アプリケーション・サーバ１５０は、グリッド・ホスト１２０内の１つまたは複数のＷｅｂサービス・コンテナにある個々のＷｅｂサービスをインスタンス化するように構成された、Ｗｅｂサービス・エンジン１６０とリンクすることができる。このＷｅｂサービス・インスタンスは、グリッド・ホスト１２０のリソース１３０にアクセスすることができる。当分野の技術者であれば、リソース１３０の集合をサービス提供インフラストラクチャの基礎とみなすことができることを理解されよう。そのため、リソース１３０は、サーバ・コンピュー

ティングのデバイスおよびプロセス、ストレージ・システム、ならびに通信およびコンピューティングの帯域幅を含むことができる。

【００２２】

重要なことには、グリッド・サービス・メカニズム１７０を各グリッド・ホスト１２０内に配置することができる。グリッド・サービス・メカニズム１７０は、ＯＧＳＡによって定義され、たとえばグローバス・プロジェクトのグローバス・ツールキット機能（非特許文献３を参照）に従って指定されたような、グリッド・サービス・インターフェースを実施することができる。当分野で周知であるように、ＯＧＳＡに準拠したグリッド・サービス・インターフェースは、以下のインターフェースおよび挙動を含むことができる。

１．Ｗｅｂサービスの作成（ファクトリ）

２．グローバル・ネーミング（グリッド・サービス処理）および参照（グリッド・サービス参照）

３．寿命管理

４．登録および発見

５．認証

６．通知

７．並行処理

８．管理の容易性

この点に関して、グリッド・サービス・メカニズム１７０は、「ファクトリ作成サービス」を使用して、選択されたＷｅｂサービスのインスタンスのクローンを新しいかまたは既存のアプリケーション・コンテナ内に作成することのできる、ファクトリ・インターフェースを含むことができる。

【００２３】

重要なことに、グリッド・サービス・メカニズム１７０は、要求されたＷｅｂサービスのクローン・インスタンスを、１つまたは複数のリモート・グリッド・ホスト１２０にまたがってインスタンス化することができる。特に、グリッド・アーキテクチャの目的と矛盾することなく、個々のリモート・グリッド・ホスト１２０により経験する処理負荷が受入れ可能な容量または事前に指定された容量を超える場合、選択されたＷｅｂサービスの新しいインスタンスをホストするために、他の個々のリモート・グリッド・ホスト１２０を選択することができる。いかなる場合でも、ルーティング・プロセス２００Ｂは、指定されたＷｅｂサービス内で処理するサービス要求の受取りに応答して、指定されたＷｅｂサービスの特定のインスタンスとは関係なく、サービス要求を処理するためにグリッド・ホスト１２０内の特定のサービス・インスタンスを選択することができる。

【００２４】

重要なことに、特定のサービス・インスタンスを選択する場合、特定のサービス・インスタンスのグリッド・ホスト１２０のサービス提供インフラストラクチャに関連付けられたリソース１３０を考慮の対象とすることができる。より詳細には、グリッド・ホスト１２０のリソース可用性を、サービス要求のリソース要件と突き合せることができる。リソースの突合せに着手するために、モニタ・プロセス２００Ａは、グリッド・ホスト１２０内で処理される各トランザクションについて、グリッド・ホスト１２０内でのリソース１３０の使用をモニタすることが可能であり、それによってトランザクションのリソース要件および消費モデルを確立することができる。各トランザクションについて確立されたモデルは、モデル・テーブル２００Ｃにストアすることができる。

【００２５】

その後、ルータは、ルーティング・プロセス時に考慮中の各グリッド・ホスト１２０について、リソース・ベクトルを確立することができる。リソース・ベクトルは、グリッド・ホスト１２０のサービス提供インフラストラクチャの基礎を形成する、個々のリソース１３０のスカラ値を含むことができる。その例には、使用可能な処理帯域幅、使用可能な通信帯域幅、ストレージ・タイプ、容量、および応答性、サーバ・タイプなどを含むことができる。グリッド・ホストのサービス提供インフラストラクチャについて確立された各

リソース・ベクトルは、ベクトル・テーブル２００Ｄにストアすることができる。さらに、ベクトルについてのコスト要素を計算することが可能であり、その結果、ベクトル・テーブル２００Ｄ内の個々のベクトルをスカラ様式で互いに比較することができる。

【００２６】

ルーティング・プロセス２００Ｂでサービス要求が受け取られると、ルーティング・プロセス２００Ｂは、サービス要求に関連付けられたトランザクション・タイプを識別することができる。このトランザクション・タイプに基づいて、モデル・テーブル２００Ｃからトランザクション・タイプのモデルを検索し、受け取ったサービス要求を処理できる使用可能なサービス・インスタンス、または受け取ったサービス要求を処理できるサービス・インスタンスをインスタンス化するための機能のいずれかを有するグリッド・ホスト１２０に関連付けられた、ベクトル・テーブル２００Ｄ内のリソース・ベクトルと突き合せることができる。この点に関して、要求を処理するために適切なグリッド・ホスト１２０を選択するように、最適アルゴリズムを適用することができる。

【００２７】

図２は、図１のグリッド内で、要求されたサービスのリソース要件に最も適したリソースを有するサービス提供インフラストラクチャ内のサービス・ホストに、サービス要求をルーティングするためのプロセスを示す流れ図である。始めに、ブロック２１０でグリッド・サービス要求を受け取ることができる。ブロック２２０では、サービス・タイプを識別することができる。ブロック２３０では、それぞれのリソース・ベクトルを確立するために、要求されたサービス・タイプのサービス・インスタンスをホストするように構成された使用可能なグリッド・ホストのリソースを照会することができる。さらに、意思決定ブロック２３０では、サービス・タイプについてのモデルが計算されたかどうかを判定することができる。

【００２８】

意思決定ブロック２４０で、識別されたサービス・タイプについてのモデルが見つからない場合、ブロック２８０で、待ち行列長さまたはスカラ・パフォーマンスに関して最高の可用性を示す、要求されたサービス・タイプのサービス・インスタンスをホストするように構成されたグリッド・ホストを選択することができる。そうでない場合は、ブロック２５０で、サービス・タイプについてのリソース・モデルを検索し、ブロック２６０で、要求されたサービス・タイプのサービス・インスタンスをホストすることができるグリッド・ホスト・セットのモデルおよびリソース・ベクトルに、最適分析を適用することができる。ブロック２６０の最適分析に基づいて、ブロック２７０では、サービス要求を特定のグリッド・ホスト内にあるサービス・インスタンスにルーティングすることができる。

【００２９】

本発明は、ハードウェア、ソフトウェア、またはハードウェアとソフトウェアの組合せにおいて実現可能である。本発明の方法およびシステムの実施は、１つのコンピュータ・システム内での中央集中方式で、あるいは様々な要素がいくつかの相互接続されたコンピュータ・システムにまたがって拡散された分散方式で、実現可能である。本明細書に記載された方法を実施するように適合されたどのような種類のコンピュータ・システムまたは他の装置も、本明細書に記載された機能の実行に好適である。

【００３０】

ハードウェアおよびソフトウェアの典型的な組合せは、ロードされ実行されると、本明細書に記載された方法を実行するようにコンピュータ・システムを制御する、コンピュータ・プログラムを備えた汎用コンピュータ・システムであってよい。本発明は、本明細書に記載された方法の実施をイネーブルにするすべての機能を含み、コンピュータ・システムにロードされるとこれらの方法を実行できる、コンピュータ・プログラム製品に組み込むことも可能である。

【００３１】

本コンテキストでのコンピュータ・プログラムまたはアプリケーションとは、情報処理機能を有するシステムに、直接あるいは、ａ）他の言語、コード、または表記法への変換

およびｂ）異なる材料形式での再生成のいずれかまたは両方を行った後、のどちらかに、特定の機能を実行させるように意図された、任意の言語、コード、または表記法での命令セットの任意の表現を意味するものである。重要なことに、本発明は、その精神または不可欠な属性を逸脱することなく他の特有の形式で具体化することが可能であり、したがって、本発明の範囲を示すものとしては、前述の明細書ではなく、添付の特許請求の範囲を参照されたい。

【００３２】

まとめとして、本発明の構成に関して以下の事項を開示する。

【００３３】

（１）サービス要求ルーティング・システムであって、

リソース・モデルをストアするように構成されたモデル・テーブルと、

前記モデル・テーブルに結合され、さらに、サービス提供インフラストラクチャ内のリソース消費をモデル化するように、および前記モデル化されたリソース消費を前記モデル・テーブル内にストアするようにもプログラミングされたモニタと、

前記モデル・テーブルに結合され、各サービス要求を、前記サービス提供インフラストラクチャのリソース構成要素と前記サービス要求のリソース・モデルとの突合せに基づいて、サービス提供インフラストラクチャを有する関連付けられたサービス・ホスト内に配置された対応するサービス・インスタンスにルーティングするようにプログラミングされたルータと

を含む、サービス要求ルーティング・システム。

（２）前記関連付けられたサービス・ホストはグリッド・コンピューティング・システム内にグリッド・ホストを含む、請求項１に記載のサービス要求ルーティング・システム。

（３）前記モデル・テーブル内の各リソース・モデルは時系列モデルである、請求項１に記載のサービス要求ルーティング・システム。

（４）前記リソース構成要素は前記サービス提供インフラストラクチャに対応するリソース・ベクトルを形成する、請求項１に記載のサービス要求ルーティング・システム。

（５）前記リソース・ベクトル内の各リソース構成要素は、サーバ・タイプ、帯域幅、およびストレージシステム・タイプからなるグループから選択されたリソースを含む、請求項４に記載のサービス要求ルーティング・システム。

（６）個々のリソース・ベクトル間の相対的コストを決定するために、各リソース・ベクトルのスカラ・コストを比較するようにプログラミングされたコンパレータをさらに含む、請求項４に記載のサービス要求ルーティング・システム。

（７）サービス要求をサービス提供インフラストラクチャのサービス・インスタンスにルーティングするための方法であって、

サービス要求を受け取るステップと、

対応するサービス提供インフラストラクチャを有する少なくとも２つのサービス・ホストについて、リソース・ベクトルを計算するステップと、

前記サービス要求についてリソース・モデルを検索するステップと、

最適のリソース・ベクトルを識別するために、前記検索されたリソース・モデルを前記リソース・ベクトルそれぞれと突き合せるステップと、

前記サービス要求を、前記識別された最適のリソース・ベクトルに関連付けられた選択されたサービス・ホストにルーティングするステップと

を含む、方法。

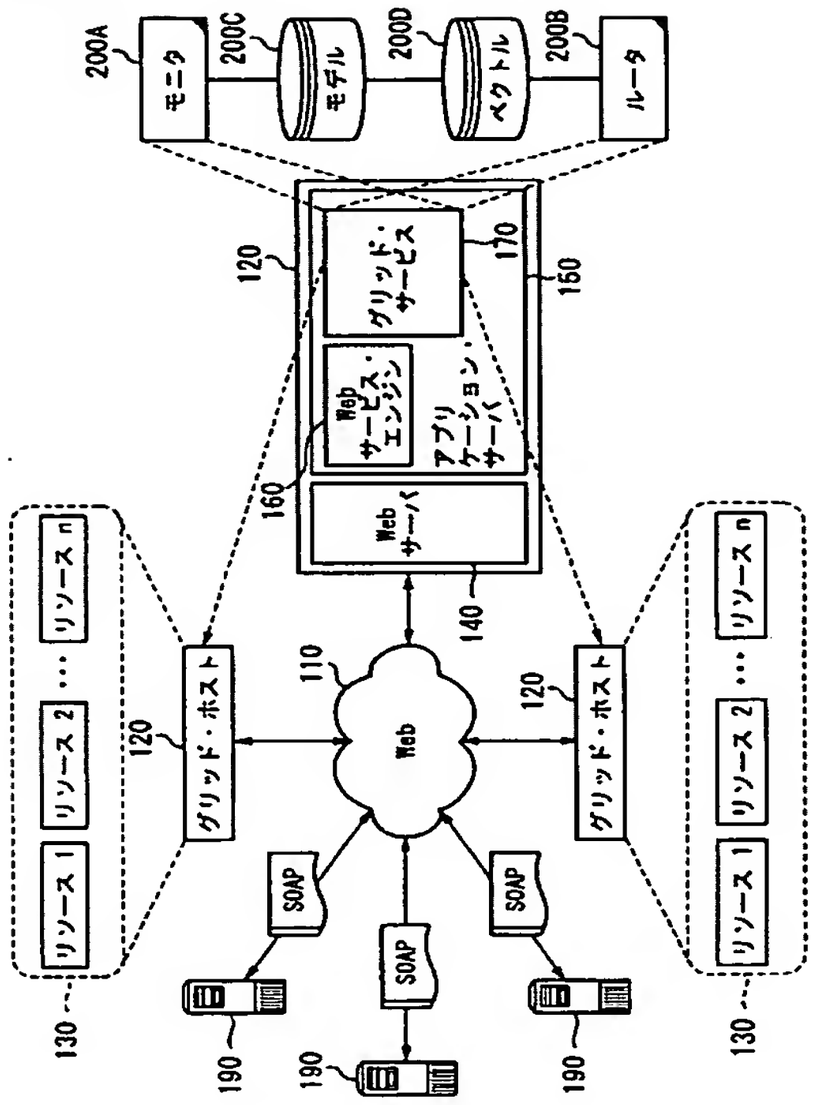
（８）前記計算するステップは、前記リソース・ベクトルのそれぞれについて、サーバ・タイプ、サーバ・パフォーマンス、サーバ容量、処理帯域幅、通信帯域幅、ストレージ・タイプ、ストレージ容量、およびストレージ・パフォーマンスからなるグループから選択された、少なくとも２つのスカラ・リソース構成要素を計算するステップを含む、請求項７に記載の方法。

（９）前記計算するステップは、前記リソース・ベクトルのそれぞれについてスカラ・コ

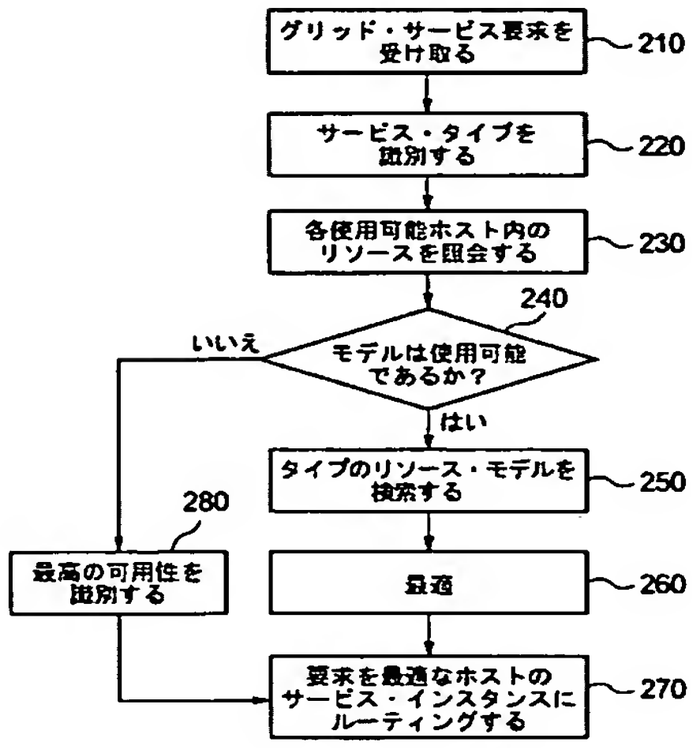
ストを計算するステップをさらに含む、請求項７に記載の方法。
（１０）受け取ったサービス要求の処理をモニタするステップと、
前記処理中に消費される、サービス・ホスト内の個々のリソース構成要素を識別するステップと、
前記サービス・ホスト内の前記識別された個々のリソース構成要素に基づいて、前記サービス要求についてのリソース・モデルを生成するステップと
をさらに含む、請求項７に記載の方法。
（１１）前記生成するステップは、前記識別された個々のリソース構成要素に基づいて、前記サービス要求のそれぞれについて時系列モデルを計算するステップを含む、請求項１０に記載の方法。
（１２）サービス要求をサービス提供インフラストラクチャのサービス・インスタンスにルーティングするために、コンピュータ・プログラムをストアしたマシン読取り可能ストレージであって、前記コンピュータ・プログラムは、
サービス要求を受け取るステップと、
対応するサービス提供インフラストラクチャを有する少なくとも２つのサービス・ホストについて、リソース・ベクトルを計算するステップと、
前記サービス要求についてリソース・モデルを検索するステップと、
最適のリソース・ベクトルを識別するために、前記検索されたリソース・モデルを前記リソース・ベクトルのそれぞれと突き合せるステップと、
前記サービス要求を、前記識別された最適のリソース・ベクトルに関連付けられた選択されたサービス・ホストにルーティングするステップと
を前記マシンに実行させるための命令のルーチン・セットを含む、マシン読取り可能記憶装置。
（１３）前記計算するステップは、前記リソース・ベクトルのそれぞれについて、サーバ・タイプ、サーバ・パフォーマンス、サーバ容量、処理帯域幅、通信帯域幅、ストレージ・タイプ、ストレージ容量、およびストレージ・パフォーマンスからなるグループから選択された、少なくとも２つのスカラ・リソース構成要素を計算するステップを含む、請求項１２に記載のマシン読取り可能記憶装置。
（１４） 前記計算するステップは、前記リソース・ベクトルそれぞれについてスカラ・コストを計算するステップをさらに含む、請求項１２に記載のマシン読取り可能記憶装置。
（１５）受け取ったサービス要求の処理をモニタするステップと、
前記処理中に消費される、サービス・ホスト内の個々のリソース構成要素を識別するステップと、
前記サービス・ホスト内の前記識別された個々のリソース構成要素に基づいて、前記サービス要求についてのリソース・モデルを生成するステップと
をさらに含む、請求項１２に記載のマシン読取り可能記憶装置。
（１６）前記生成するステップは、前記識別された個々のリソース構成要素に基づいて、前記サービス要求のそれぞれについて時系列モデルを計算するステップを含む、請求項１５に記載のマシン読取り可能記憶装置。

【図面の簡単な説明】
【００３４】
【図１】本発明に従い、要求されたサービスのリソース要件に最も適したリソースを有するサービス提供インフラストラクチャ内のサービス・ホストに、サービス要求をルーティングするように構成されたサービス・グリッドのブロック図である。
【図２】図１のグリッド内で、要求されたサービスのリソース要件に最も適したリソースを有するサービス提供インフラストラクチャ内のサービス・ホストに、サービス要求をルーティングするためのプロセスを示す流れ図である。

【図１】


200A
モニタ
200C
モデル
200D
ベクトル
200B
ルータ
120
グリッド・サービス
170
150
Web サービス・エンジン
アプリケーション・サーバ
160
Web サーバ
140
110
Web
120
グリッド・ホスト
グリッド・ホスト
SOAP
SOAP
SOAP
190
190
190
130
130
リソース n
リソース 2
リソース 1
リソース n
リソース 2
リソース 1

【図２】


グリッド・サービス要求を受け取る
210
サービス・タイプを識別する
220
各使用可能ホスト内のリソースを照会する
230
240
モデルは使用可能であるか？
いいえ
はい
タイプのリソース・モデルを検索する
250
280
最高の可用性を識別する
最適
260
要求を最適なホストのサービス・インスタンスにルーティングする
270

フロントページの続き

(72)発明者　ロナルド・ピー・ドイル
アメリカ合衆国２７６１５　ノースカロライナ州ラーレー　アボカド・サークル　１００００
(72)発明者　デビッド・ルイス・カミンスキー
アメリカ合衆国２７５１４　ノースカロライナ州チャペル・ヒル　コービン・ヒル・サークル　１０３

ドターム(参考) 5B045 BB11 BB28 BB42 GG02
5B098 AA10 GA01 GD02 GD14
5K030 HC01 JT06 KA05 KA07 LB05

【要約の続き】